

《Web数据挖掘》

web scraper (4)

教师: 林志良

邮箱: linzhl@nfu.edu.cn

个人网站: www.zhilianglin.com

目录

- grouped选择器
- element click选择器

grouped选择器

什么时候可以使用grouped选择器？

- 如果目标抓取内容是**多组相似的文本**，可以使用grouped选择器。grouped选择器会将这些文本**打包在一起**，并以json格式存储。
- 应用场景示例：**抓取电子商务网站商品的尺寸、型号等信息。

Data Preview

name	price	color	size
优衣库男装女装宽松直筒牛仔裤 水洗产品阔感裤长裤常规款 470542	299	[{"color":"01 乳白色"}, {"color":"08 深灰色"}, {"color":"64 湖蓝色"}, {"color":"68 深蓝色"}]	[{"size":"160/68A"}, {"size":"160/70A"}, {"size":"165/74A"}, {"size":"165/76A"}, {"size":"170/78A"}, {"size":"170/82A"}, {"size":"175/84A"}, {"size":"175/86A"}, {"size":"175/88A"}, {"size":"180/90B"}, {"size":"180/96B"}, {"size":"185/100B"}, {"size":"185/106C"}, {"size":"185/112C"}, {"size":"190/116C"}, {"size":"190/122C"}]

grouped选择器

实战：抓取淘宝网某商品基本信息

- 网址：

https://detail.tmall.com/item.htm?abbucket=5&id=810826299002&ns=1&priceTId=214782ba17275133795162677e17ff&skuid=5672200302204&spm=a21n57.1.item.3.1188523cqappPW&utparam=%7B%22aplus_abtest%22%3A%220156d5b363c615f674ef17caa60b7bf6%22%7D&xxc=taobaoSearch (需要登录)

- 要求：抓取商品名、价格、颜色、尺码

- 提示：抓取多项内容，我们需要先使用element选择器，再在其下层级使用grouped选择器和text选择器完成抓取任务

grouped选择器

实战：抓取淘宝网某商品基本信息

优衣库男装女装宽松直筒牛仔裤水洗产品阔感裤长裤常
规款470542

已售 6000+ | 可开发票

国庆狂欢 热卖中 下单立抢

¥299

会员券满500减50 同地址满1件包邮 购买得积分

活 动：会员专享 50元券，满500元可用

配 送：多仓发货 至 广州市从化区 ~
快递：免运费 承诺48小时内发货，晚发必赔

保 障：假一赔四 退货宝 极速退款 ~

颜 色：
01 乳白色
08 深灰色
64 湖蓝色
68 深蓝色

尺 码：
160/68A
160/70A
165/74A
165/76A
170/78A
170/82A
175/84A
175/86A
175/88A
180/90B
180/96B
185/100B
185/106C
185/112C
190/116C
190/122C

grouped选择器

实战：抓取淘宝网某商品基本信息

The screenshot shows the CSS Selector Gadget interface. At the top, there's a preview of a product page with a color selection section. The section has a label '颜色:' followed by four color swatches with corresponding labels: '01 乳白色', '08 深灰色', '64 湖蓝色', and '68 深蓝色'. A red arrow points from the 'Type' dropdown in the configuration panel down to the 'color' swatch in the preview.

Id	color
Type	Grouped
Selector	Select
Attribute name	Click to view available attributes
Parent Selectors	_root element

多选多个目标元素
一般不填写，如果填写会抓取元素属性内容

Save selector Cancel

grouped选择器

练习：抓取淘宝网iphone 16 Pro Max不同配置选择

- 网址：

https://detail.tmall.com/item.htm?abbucket=5&id=833003275432&ns=1&priceTId=214780a417275279875852694e626a&skuid=5581805713350&spm=a21n57.1.item.98.79ef523cXvKA0b&utparam=%7B%22aplus_abtest%22:%225c340418ce41cd00f5ae5c51432c1aa8%22%7D&xxc=taobaoSearch (需要登录)

- 要求：抓取机身颜色、存储容量

grouped选择器

练习：抓取淘宝网iphone 16 Pro Max不同配置选择

Apple/苹果 iPhone 16 Pro Max

可开发票

¥9999

优惠： 24期免息，购买得积分

保障： 极速退款，七天无理由退换

配送： 多仓发货 至 广州市 从化区 ▾
快递: 免运费 付款后31天内发货

版本：

iPhone 16 Pro

iPhone 16 Pro Max

机身颜色：

沙漠色钛金属

原色钛金属

白色钛金属

黑色钛金属

存储容量：

256GB

512GB

1TB

保障服务： AppleCare+ 服务计划（不限次意外损坏保修服务）

购买 1 年 ¥899

购买 2 年 ¥1599

数量：

-

1

+

(限购2件)

element click选择器

什么时候需要使用element click选择器

- 当需要**模拟鼠标点击**网页的时候，可以考虑使用element click选择器（如有的网站需要点击【阅读更多】才会加载更多内容出来；一些网页点击特定区域后会加载新的内容出来）
- 注意：** element click选择器是具有鼠标点击功能的element选择器
(selector需要选择两个)

element click选择器

element click选择器操作界面

Sitemaps Sitemap wallstreetcn ▾ Create new sitemap ▾

Id click

Type Element click 选择要抓取的内容区域（这里还是结构体，不包含信息）

Selector Select Element preview Data preview

Click selector Select Element preview div.more-link 选择需要进行鼠标点击的区域

Click type Click once (pagination, tabs) 点击一次/多少次

Click element uniqueness unique CSS Selector 根据什么（文本/html/CSS）判断点击区域的唯一性

Multiple

Discard initial elements Never discard 是否要丢弃点击鼠标之前抓取到的内容

Delay (ms) 2000

Parent Selectors _root
click

element click选择器

实战：抓取华尔街见闻首页新闻（点击【加载更多】）

- 网址：<https://wallstreetcn.com/>
- 要求：抓取点击【加载更多】后的新闻标题



以色列空军袭击黎真主党总部，油价反弹，穆迪下调以色列评级
以军正在对黎巴嫩南部真主党的军事目标展开新一轮袭击。以色列空军袭击了位于黎巴嫩首都贝鲁特的黎真主党总部。以色列国防...
何浩 | 9小时前



高盛顶级交易员回应投资者最关切的问题：全球股市，追涨还是不追？
高盛对冲基金研究主管Tony Pasquariello在最新报告中表示，他依然相信市场的主要趋势是向上的，而且肯定不打算站在央行“火炮”...
何浩 | 10小时前



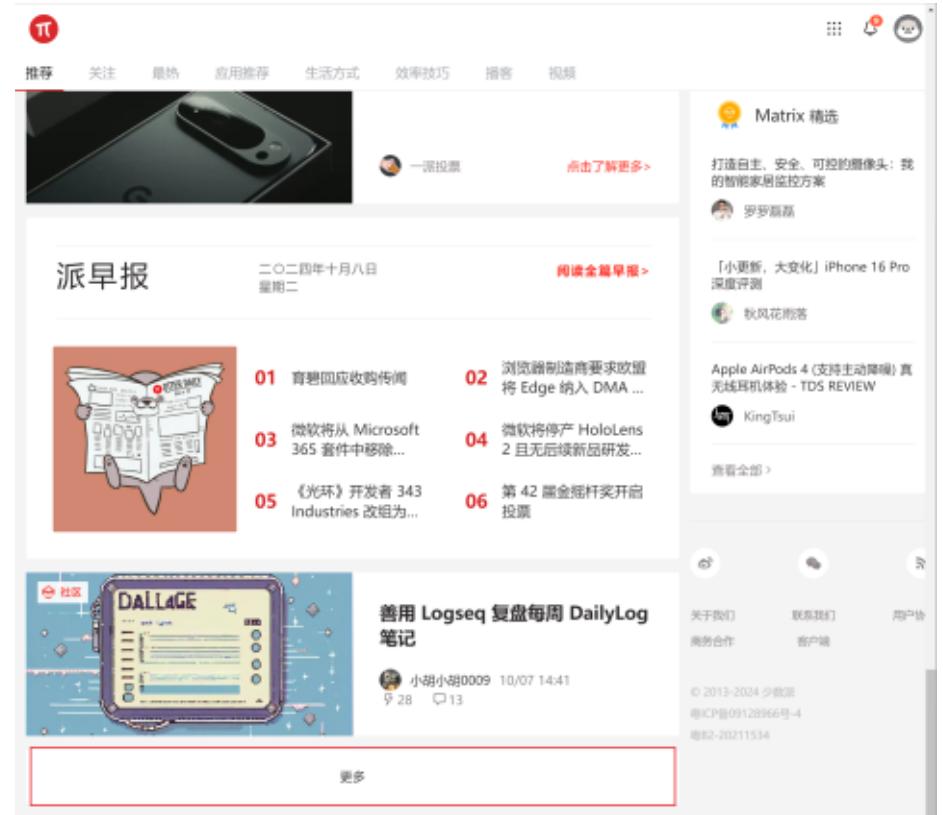
美国GDP上修“玄机”？商务部一个月让国民储蓄增加5000亿
GDP得到“意外”提振源于美国商务部修改了个人收入和支出，税后个人可支配收入上调 3.8%，支出上调幅度不到收入的一半。由此计...
李丹 | 10小时前

加载更多 >

element click选择器

练习：抓取少数派首页文章标题（点击【加载更多】）

- 网址：<https://sspai.com/>
- 要求：抓取点击一次【加载更多】后的文章标题



element click选择器

实战：抓取淘宝网iphone 16 Pro Max不同存储容量的价格信息

- 网址：

https://detail.tmall.com/item.htm?abbucket=5&id=833003275432&ns=1&priceTId=214780a417275279875852694e626a&skuid=5581805713350&spm=a21n57.1.item.98.79ef523cXvKA0b&utparam=%7B%22aplus_abtest%22:%225c340418ce41cd00f5ae5c51432c1aa8%22%7D&xxc=taobaoSearch (需要登录)

- 要求：抓取存储容量、价格

element click选择器

练习：抓取淘宝网小米14手机不同存储容量的价格信息

- 网址：

https://detail.tmall.com/item.htm?abbucket=5&id=742106313947&ns=1&priceTId=2150462c17275388632945664ec9a7&skuid=5136864796771&spm=a21n57.1.item.194.79ef523cXvKA0b&utparam=%7B%22aplus_abtest%22%3A%22baf47227d49735b3394084fbfa2b67b%22%7D&xxc=taobaoSearch (需要登录)

- 要求：抓取存储容量、价格

参考资料

- 大数据Anniee酱的个人空间-大数据Anniee酱个人主页-哔哩哔哩视频

<https://space.bilibili.com/52180557/channel/seriesdetail?sid=665534>

- Web scraper documentation

<https://www.webscraper.io/documentation>

- 21堂不写代码的信息掘金课 爬虫

- Html, CSS 等前端知识请参考：<https://www.runoob.com/>

参考资料

- 卤蛋实验室 的文章

<https://sspai.com/u/skychx/posts>

谢谢！